

Projekt Speech2Text

Automatizovaná transkripce

Martin Rechterik - Národní archiv
- Západočeská univerzita v Plzni



Archivy, knihovny, muzea v
digitálním světě 2025 (26. ročník)



Select * from “Me”

- magistr, historické vědy, UPOL
- metodik v Národním archivu
- doktorské studium na FAV, Geomatika ZČU
- člen [RDB AIG](#) při [DILCIS Board](#) (DLM Forum)
- člen pracovní skupiny pro databáze pod patronací [NARA](#)
- vedoucí Pracovního týmu pro národní infrastrukturu pro prostorové informace se zaměřením na archivaci prostorových dat (PT NIPI_ARCHI)
- martin.rechtorik(at)na.gov

Schéma

- Aplikovaný transformerový/transformátorový model
- Archivní fond BBC - česká redakce, Londýn, Praha
- Výsledky měření transkripce pomocí CER a WER
- Procesní workflow pro tvorbu SIP

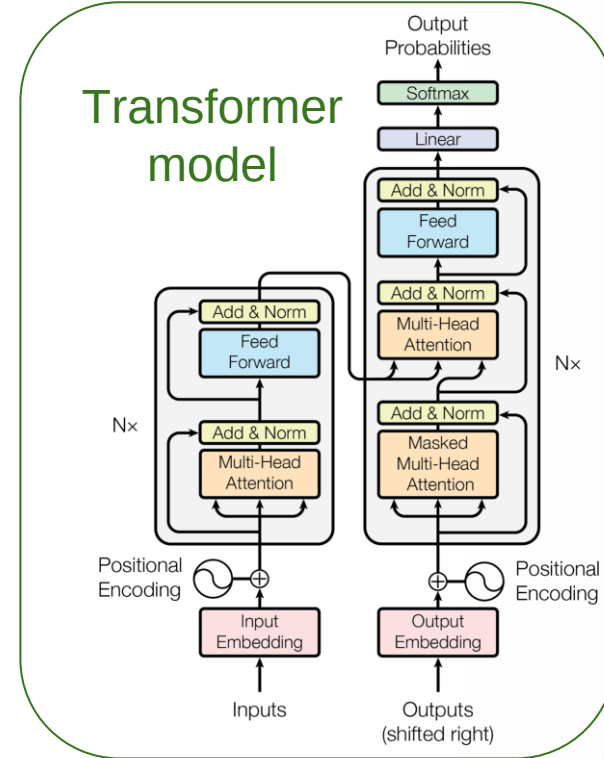
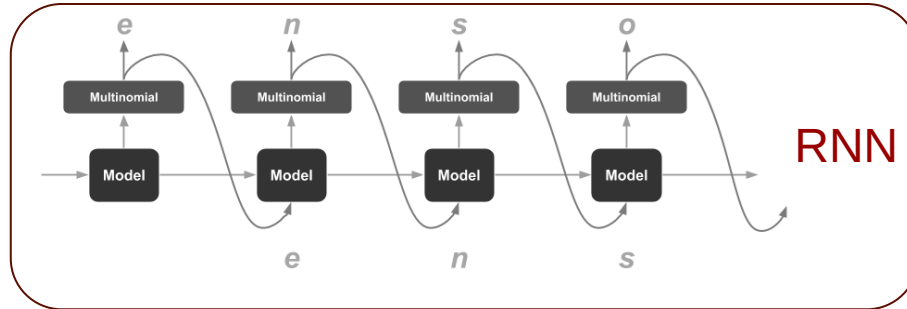
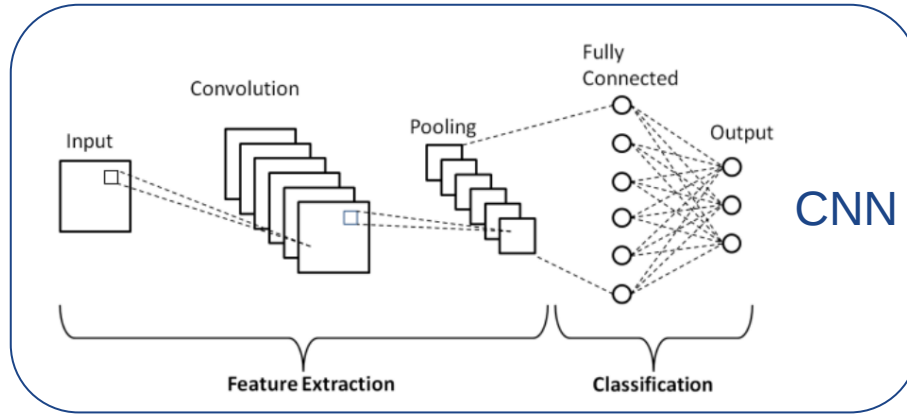
Speech2text, Transformátorový model

- Transformátorový model se zcela opírá o mechanismy sebeuvědomění a opomíjí rekurentní a konvoluční vrstvy.
- Sebeuvědomění používá k výpočtu reprezentací vstupních a výstupních sekvencí, což mu umožňuje zvážit důležitost různých slov ve větě.
- Architektura využívá “multi-head attention”, což umožňuje modelu společně zpracovávat informace z různých reprezentačních podprostorů na různých pozicích.
- Jelikož model nemá zabudovanou představu o pořadí sekvence, zahrnuje poziční kódování, které modelu poskytuje informace o pozici slov v sekvenci.
- Transformátor se skládá z enkodéru, který zpracovává vstup, a dekodéru, který generuje výstup. Oba jsou složeny z více vrstev neuronových sítí se sebevnímáním a předáváním.

Speech2text, Transformátorový model

- Model je trénovaný pomocí varianty stochastického gradientního sestupu a využívá techniky jako vyhlazování štítků a dropout k zlepšení generalizace.
- Transformátorový model dosahuje výborných výsledků v různých testech strojového překladu nebo přepisů a překonává předchozí modely, které se opíraly o rekurentní architektury.
- Architektura je vysoce paralelizovatelná, což umožňuje efektivnější trénování na velkých datových souborech ve srovnání s tradičními RNN.
- Zavedení Transformátoru významně ovlivnilo oblast zpracování přirozeného jazyka a vedlo k vývoji různých modelů, jako jsou BERT a GPT.

CNN, RNN, Transformátorový model



Národní archiv: BBC - česká redakce, Londýn, Praha 2001-2007

■ Základní informace o [archivní sbírce](#)

- vybráno za archiválie v roce 2013, Miroslav Kunt
- trvale uchovává Národní archiv
- zvukové nahrávky a jejich základní a nedostatečný popis v podobě strukturovaných dat (MS Excel + MS Access)
- Velikost 850,313 GBs, 14 068 audio nahrávek
- délku odhadujeme na 305 dní celkem
- BBC archiv (cca 927 nahrávek s délkou více než 7 dní)
- BBC červený archiv (13 141 nahrávek, 297 dní)
- datový formát: MP2, WAV

Vysílání v českém a slovenském jazyce na BBC

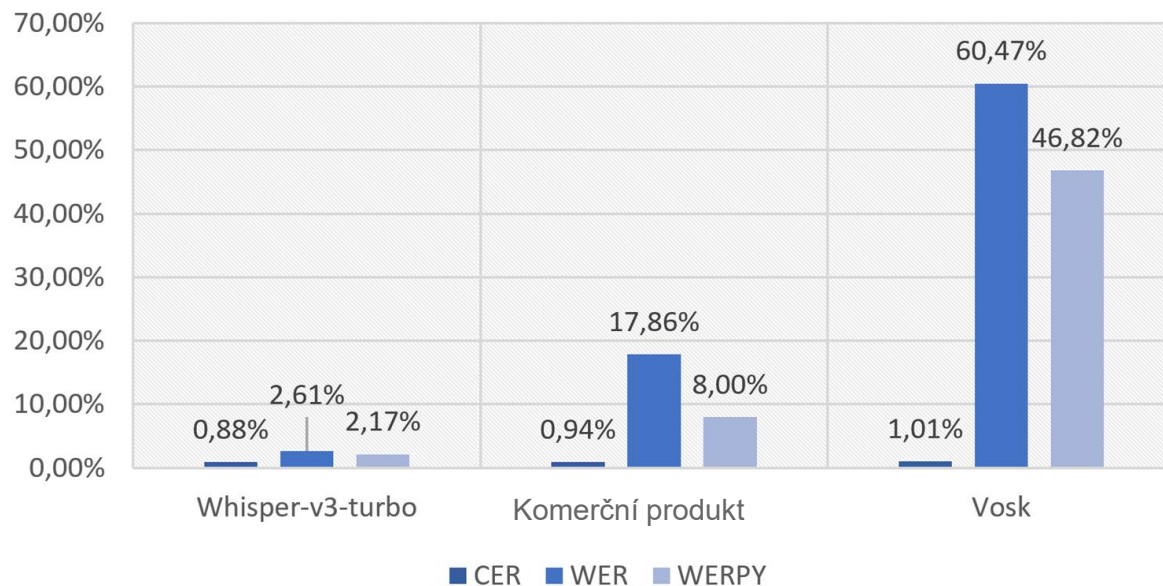
- Vysílání v českém a slovenském jazyce zahájeno 6. 9. 1939, jen několik dní po napadení Polska nacistickým Německem.
- Vysílání bylo iniciováno ze strany Labour Party, která jako první podpořila obnovení Československa po uzavření Mnichovské dohody a v důsledku německé okupace v březnu 1939.
- První proslov na BBC přednesl na podporu svých krajanů v Československu sociální demokrat Josef Kosina.
- Pod vlajkou the Overseas Service, proběhl v následujících letech velký rozvoj cizojazyčného vysílání včetně toho Československého.
- Kromě Československého vysílání probíhalo i vysílání v českém a slovenském jazyce, které bylo známé jako Hlas svobodné republiky. Jednalo se o vysílání Československé vlády v exilu reprezentované prezidentem Edvardem Benešem nebo ministrem zahraničních věcí Janem Masarykem.

Vysílání v českém a slovenském jazyce na BBC

- Československá redakce pokračovala ve své činnosti i po skončení druhé světové války.
- Během období komunistické totality až do jejího konce v roce 1989 představovalo vysílání BBC spolehlivý zdroj informací z demokratického světa.
- Mezi významné osobnosti, které se účastnily programů BBC v československém jazyce, patřili Karel Kyncl, bratr Jan Lang S.I., Bedřich Bělohlávek, Karel Brušák, Zdena Tominová, Petr Brod, Vít Kolář a mnoho dalších.
- Po sametové revoluci se česká redakce BBC stala líhní známých moderátorů a novinářů působících ve veřejnoprávních i komerčních médiích, jako jsou Václav Moravec, Martin Řezníček, Jan Bumba, Daniel Kaiser nebo Ondřej Štindl.

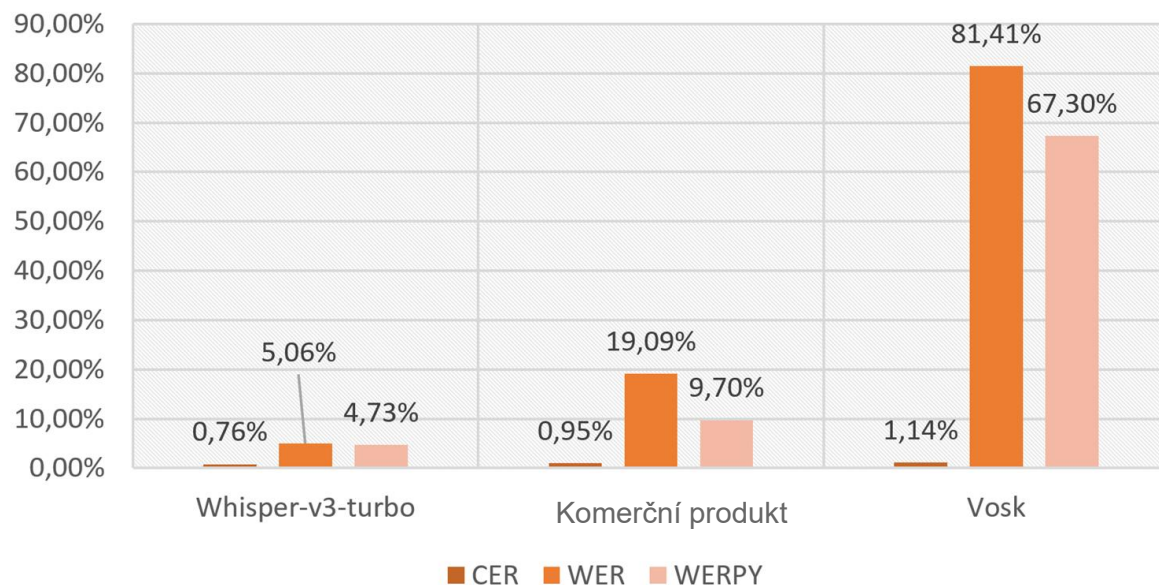
Výsledky: Nahrávka 1

Zjištěná chybovost

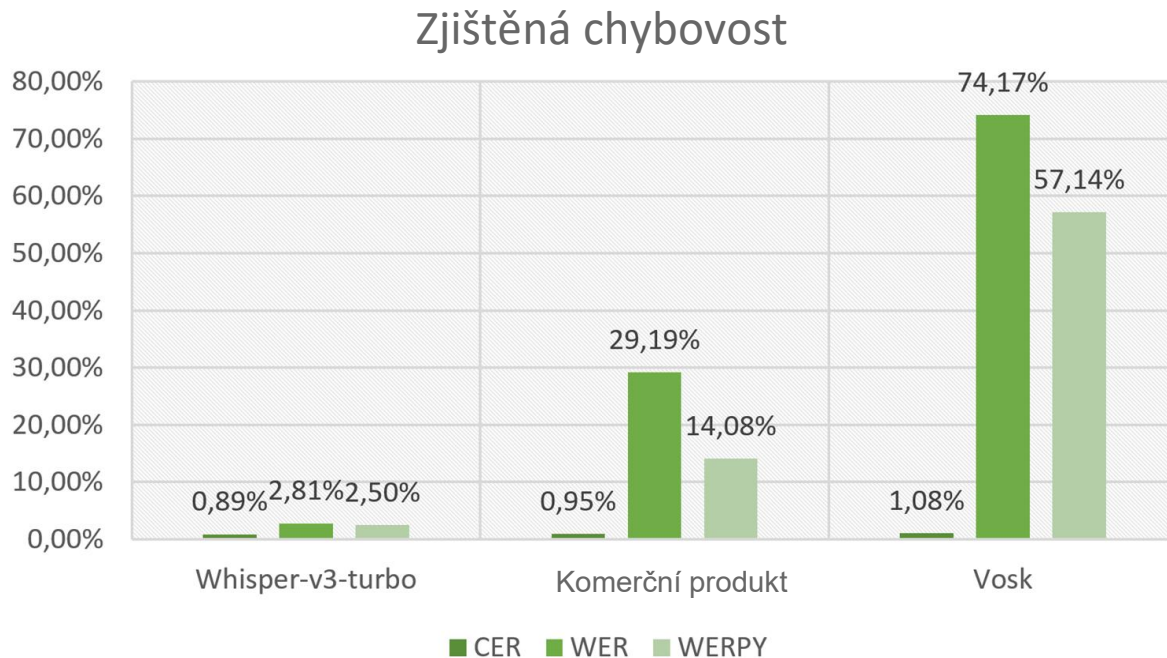


Výsledky: Nahrávka 2

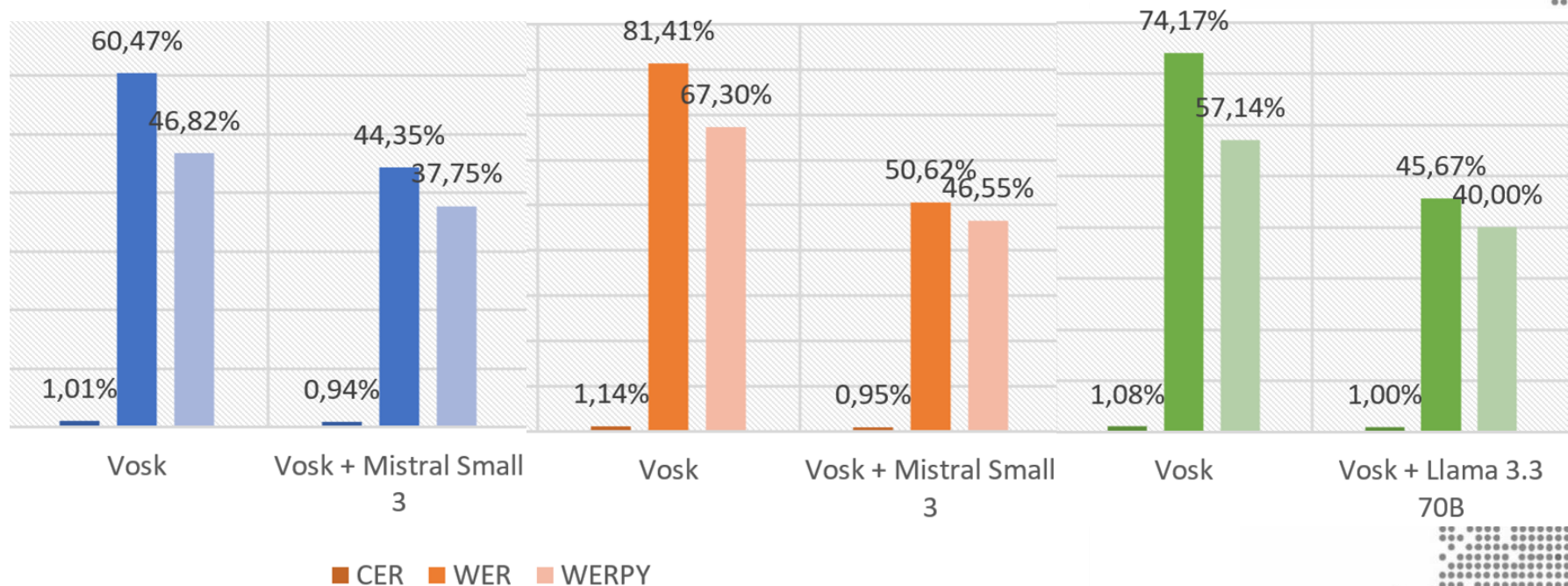
Zjištěná chybovost



Výsledky: Nahrávka 3



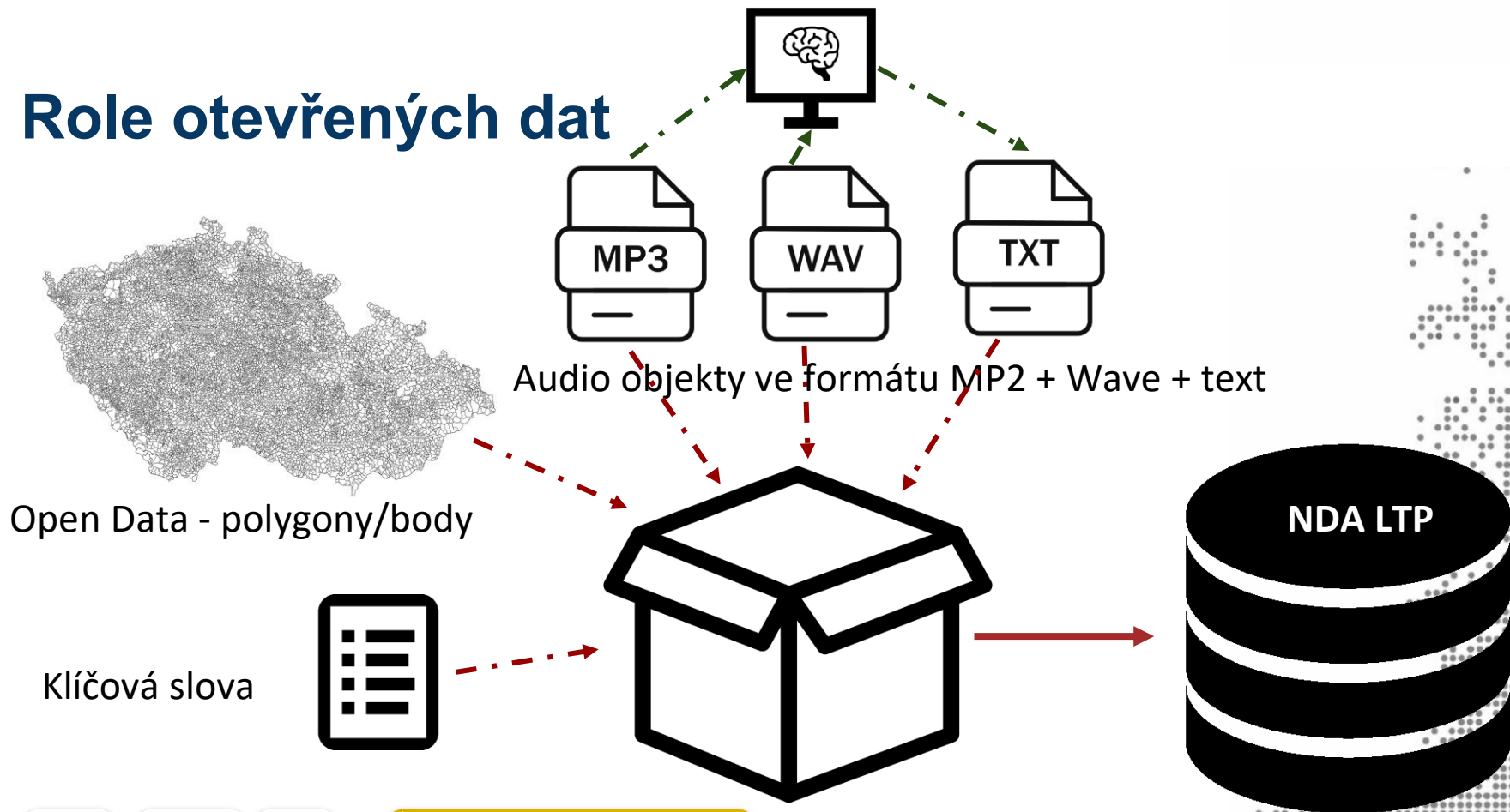
Může LLM pomoci zlepšit výsledky z Vosku



Workflow tvorby SIP

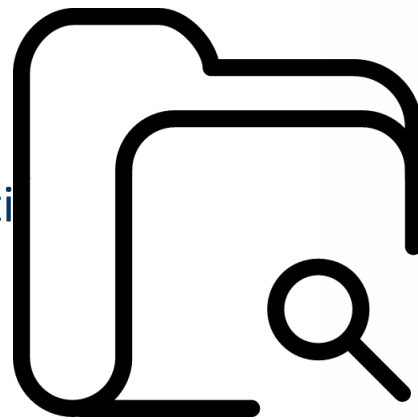
- Postup zpracování dat
 - průchod datovou strukturou, nalezení zvukových souborů, extrakce textové vrstvy
 - průnik textu s množinami klíčových slov a prostorových objektů
 - vytvoření adresářové struktury pro potřeby EARK SIP
 - tvorba metadat (Mets, EAD cz-Profile)
 - vytvoření SIP pomocí nástrojů EARK
 - nahrání dat do Národního archivního portálu
 - datový transfer NDA(LTP)
 - metadatové změny v AIPEch bude provádět pořádací software

Role otevřených dat



Možnosti přístupu

- Tradiční způsob
 - Archivní pomůcka - předpokládá LLM a RAG pro vytvoření kvalitních registů
 - Vytvoření multi modelové databáze s využitím ElasticSearch
- Moderní způsob
 - Vektorové hledání na principu příbuznosti
 - Pokročilé řazení výsledků



Shrnutí

- automatizace celého procesu je možná
- proces nutně nemusí být náročný
- spoluúčast studentů z ČVUT byla užitečná
- kromě zvuku lze extrahovat text z video souborů
- profesionální komerční nástroj je skvělý pro identifikaci mluvčích
- klíčová slova a lokality není problém dohledávat
- nástroj lze využít i pro jiné scénáře - zpřístupnění v prostředí badatelný, zápisy z porad atd.

Závěry z projektu

- nekvalitní metadata získaná od původce jsou limitující
- postrádáme implementaci LLM vč. Retrieval Augmented Generation (RAG)
- výkonná pracovní stanice vč. GPU je nezbytnost
- Faster-Whisper nepodporuje MP2 souborový formát
- Speech2Text s využitím Vosk modelu podporuje pouze jednokanálový WAV zvukový formát, ale využitelný je pouze pro nahrávky v anglickém jazyce

Reference

- ALPAYDIN, Ethem. *Machine learning*. Revised and updated edition. The MIT Press essential knowledge series. Cambridge, Massachusetts: MIT Press, [2021]. ISBN 978-0-262-54252-4.
- DRORI, Iddo. *The science of deep learning*. Cambridge: Cambridge University Press, 2023. ISBN 978-1-108-83508-4.
- FIŠER, Marek, KYKAL, T.: Digitální studovna Ministerstva obrany ČR, Využití technologií na pokročilou indexaci obsahu historických dokumentů. In Soudobé dějiny 2/2004, Available from: <https://doi.org/10.51134/sod.2024.027> [accessed 2025-04-29].
- CHOLLET, François. *Deep learning v jazyku Python: knihovny Keras, Tensorflow*. Knihovna programátora. Praha: Grada Publishing, 2019. ISBN 978-80-247-3100-1.
- KELLEHER, John D. *Deep learning*. The MIT Press essential knowledge series. Cambridge, Massachusetts: The MIT Press, [2019]. ISBN 978-0-262-53755-1.
- KOCOUREK, Milan. *Volá Londýn: historie českého a slovenského vysílání BBC : (1939-2006)*. Praha: Ottovo nakladatelství, c2013. ISBN 978-80-7451-318-3.
- RASMUSSEN, Casper Hvenegaard; RYDBECK, Kerstin a LARSEN, Håkon (ed.). *Libraries, archives, and museums in transition: changes, challenges, and convergence in a Scandinavian perspective*. Online. London: New York, 2023. ISBN 978-1-032-03364-8. Available from: <https://www.taylorfrancis.com/books/oa-edit/10.4324/9781003188834/libraries-archives-museums-transition-casper-hvenegaard-rasmussen-kerstin-rydbeck-h%C3%A5kon-larsen>. [accessed 2025-04-29]
- SHAH, Chirag. *A hands-on introduction to machine learning*. Cambridge: Cambridge University Press, 2023. ISBN 978-1-009-12330-3
- Text generation with an RNN, accessed from https://www.tensorflow.org/text/tutorials/text_generation
- Van Hiep Phung, Eun Joo Rhee A High-Accuracy Model Average Ensemble of Convolutional Neural Networks for Classification of Cloud Image Patches on Small Datasets, available from <http://dx.doi.org/10.3390/app9214500> [accessed 2025-04-29]
- VASWANI, Ashish; SHAZEER, Noam; PARMAR, Niki; USZKOREIT, Jakob; JONES, Llion; GOMEZ, Aidan N.; KAISER Lukasz; POLOSUKHIN, Illia. *Attention is all you need*. Cornell University 2023. Available from: <https://doi.org/10.48550/arXiv.1706.03762>. [accessed 2025-02-20].

Projekt Speech2Text

Automatizovaná transkripce

Martin Rehtorik - Národní archiv
- Západočeská univerzita v Plzni



Archivy, knihovny, muzea v
digitálním světě 2025 (26. ročník)



DEPARTMENT
OF GEOMATICS

University of West Bohemia in Pilsen, Faculty of Applied Sciences, Department of Geomatics. Author was supported by Project: SGS-2025-008: Využití matematiky a informatiky v geomatice VII/Application of Mathematics and Informatics in Geomatics VI.